EECS 349 (Machine Learning) Homework 3

WHAT TO HAND IN

You are to submit the following things for this homework:

1. A **PDF** document containing answers to the homework questions.
2. The **WELL COMMENTED** source code for all software you write.

   NOTE: If your code is not well commented do not expect full points. Every function should have comments associated with it that specify what the input data format is, what the output data format is and what the function does. Any tricky bits in the code should have a comment explaining them

HOW TO HAND IT IN

To submit your assignment:

1. Compress all of the files specified into a .zip file.
2. Name the file in the following manner, firstname_lastname_hw3.zip. For example, Bryan_Pardo_hw3.zip.
3. Submit this .zip file via Canvas.

DUE DATE: the start of class as specified in the course calendar.

1) Comparing kernels on real data (3 points)

Get the LibSVM support vector machine. A link is available on the course website on the "links" page. Note, you can add LibSVM to WEKA and use it within the WEKA framework. Consult the WEKA and LibSVM documentation for how.

Examine the performance of LibSVM on the ionosphere dataset (http://archive.ics.uci.edu/ml/datasets/Ionosphere) as a function of kernel and misclassification cost, ***C***. Try kernels from the set {Linear, Polynomial, Radial Basis Function}. Try the following set of values for ***C***: { 10^-2, 10^-1, 10^0, 10, 10^2}. Which combination of kernel and cost is best on this data set? Are the differences statistically significant? Use cross-validation. You'll have to pick how many folds to get some kind of statistical significance. Support your answer with figures and statistical tests. Explain why you chose the statistical test(s) you decided to use.

On installing LibSVM: Prem says "Installing LibSVM is super easy from the developer version of weka (http://www.cs.waikato.ac.nz/ml/weka/downloading.html down to developer versions) because there's a handy GUI package manager. So it might be good to suggest version 3.7.11"

On loading Ionosphere: Ionosphere is supposed to be in the c4.5 data format, but the .names file doesn't have the typical names fields, and is instead just a verbose description of the data. A way to get around this is to add a column header row and import as a csv file. Here's another way (again from Prem) "I just went ahead and found the .arff file on this page: http://repository.seasr.org/Datasets/UCI/arff/ionosphere.arff. That imported without any issues."

2) SVM for text documents (2 points)

A canonical classification task for text documents is to classify email as spam or not spam. Describe an SVM kernel good for this purpose.

EECS 349 (Machine Learning) Homework 3

A. (1/2 point) Describe a way to encode text documents so that they may be classified using an SVM. This means either defining $\varphi(x)$, a mapping from a document $\boldsymbol{x}$ to a vector space where you can do inner products or defining a kernel $k(x,x')$ without explicitly defining $\varphi(x)$ This doesn't have to be an original approach. It should be a good approach.

B. (1/2 point) If you use a kernel function in part A, explain how you know it is a kernel. If you used a mapping function $\varphi(x)$, explain how the output is something where one can perform an inner product.

C. (1/2 point) Explain why your encoding or kernel function would be useful for the problem of telling spam from good email.

D. (1/2 point) Cite the paper for building text kernels that is most related to your approach to building either $\varphi(x)$ or a kernel function. Give full bibliographic information and a web link. Explain why this is closely related to what you propose. Did you base your answer on this paper? (it is OK if you did)

3) Speeding up a Support Vector Machine (2 points)

An unoptimized implementation of a support vector machine (SVM) will take O(N^3) to find the maximal margin separator, where N is the number of data points. Assume you have an unoptimized support vector machine that takes N^3 steps to find the maximal margin separator. Assume it uses a linear kernel (The kernel isn't central to this problem, I just want to establish the kernel doesn't add to the processing time). Assume you have 10^6 labeled data points in your data split into two classes with roughly equal numbers of points.

A. (1/2 point) We can learn the optimal separating hyperplane by simply applying the SVM to the dataset of 10^6 points. How many steps would that take? How quickly would this run to completion if you could do one billion steps per second?

B. (1/2 point) Describe a way to use the same SVM to find a good (perhaps not maximal margin, but darn close) separating hyperplane for the two classes that is (at least) 10^7 times faster than the naïve approach described in step A. This method should use the unoptimized SVM without alteration.

C. (1/2 point) Give a proof of your assertion of an (at least) 10^7 speedup with the method described in step B.

D. (1/2 point) Describe how you would mathematically (i.e. statistically) model the expected difference between the decision surface you learn with the method in step B and the decision surface you would learn with the method in step A.

4) Is it a Kernel? (1 point)

Someone has asserted to you that a Gaussian-like formula is a valid kernel, where *x, x'* are vectors of real values. They present the following formulation of that kernel.

$$k(x,x') = \exp(-\|x - x'\|^2 / 2\sigma^2)$$

Is this a valid kernel or not? Give a proof. Feel free to use the rules for kernel composition given in the lecture notes. Also, you could consult the paper on the course readings titled "Kernel Methods in Machine Learning."

EECS 349 (Machine Learning) Homework 3

5) Perceptrons (2 points)

A. (1/2 point) Is a one-layer perceptron capable of building non-linear decision surfaces? Why or why not?

B. (1/2 point) Is a multi-layer perceptron capable of non-linear decision surfaces? Why or why not?

C. (1 point) If the sigmoid function in a multi-layer perceptron were replaced with the following function, how would this affect the back propagation of error training method?

$$output = sign\left(\frac{1}{1+e^{-net}}\right)$$

$$sign(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ -1 \text{ else} \end{cases}$$

6) Restricted Boltzman Machines (1 point)

In the course calendar, there is a link to a tutorial on Deep Belief Networks. There is also a paper called "Scaling Learning Algorithms towards AI"…and there is always the Wikipedia. These resources will help you answer the following questions.

A. (1/2 point) Explain what a Restricted Boltzman Machine (RBM) is. Don't just give a sentence.

B. (1/2 point) What relationship do RBMs have to Deep Belief Networks?

7) Scaling Learning Algorithms towards AI (2 points)

Read Scaling Learning Algorithms towards AI (available on the course calendar). Then answer the following questions. Note, a one sentence answer to a question is a good way to get 0 points.

A. (1/2 point) What are Kernel Machines?

B. (1/2 point) Describe two limitations of Kernel Machines.

C. (1/2 point) Give an informal description of Deep Architectures.

D. (1/2 point) How do the authors expect Deep Architectures will get around the limits you describe in B?

8) SVM VS Multilayer Perceptron (2 points)

Use the ionosphere dataset (http://archive.ics.uci.edu/ml/datasets/Ionosphere) and Weka. Compare the performance of an SVM using your best settings from question 1 to a three layer (i.e. single hidden layer) perceptron. Compare on speed and classification accuracy. Explain the experiment you ran to compare the two approaches. Describe the architecture of your neural net (how many nodes in each layer, for example). Specify your kernel and value for C for the SVM. Give classification results. Measure statistical significance. Show a figure that illustrates your results.